AirVis: Visual Analytics of Air Pollution Propagation

Zikun Deng, Di Weng, Jiahui Chen, Ren Liu, Zhibing Wang, Jie Bao, Yu Zheng, and Yingcai Wu

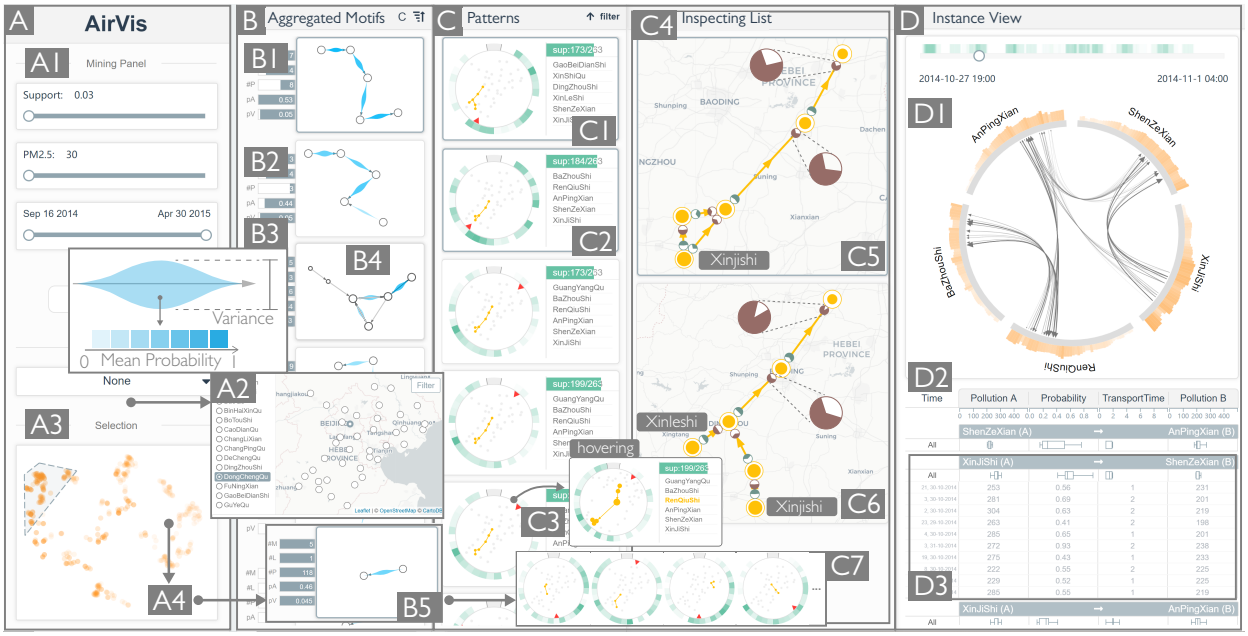

Fig. 1. The system interface of AirVis: (A) the control panel enables the interactive mining, filtering, and selection of the propagation patterns; (B) the motif view presents the extracted significant motifs with uncertainty-aware visualizations; (C) the pattern view depicts the patterns with compact pattern glyphs and pattern graphs; (D) the instance view helps users inspect the propagation instances.

Abstract— Air pollution has become a serious public health problem for many cities around the world. To find the causes of air pollution, the propagation processes of air pollutants must be studied at a large spatial scale. However, the complex and dynamic wind fields lead to highly uncertain pollutant transportation. The state-of-the-art data mining approaches cannot fully support the extensive analysis of such uncertain spatiotemporal propagation processes across multiple districts without the integration of domain knowledge. The limitation of these automated approaches motivates us to design and develop AirVis, a novel visual analytics system that assists domain experts in efficiently capturing and interpreting the uncertain propagation patterns of air pollution based on graph visualizations. Designing such a system poses three challenges: a) the extraction of propagation patterns; b) the scalability of pattern presentations; and c) the analysis of propagation processes. To address these challenges, we develop a novel pattern mining framework to model pollutant transportation and extract frequent propagation patterns efficiently from large-scale atmospheric data. Furthermore, we organize the extracted patterns hierarchically based on the minimum description length (MDL) principle and empower expert users to explore and analyze these patterns effectively on the basis of pattern topologies. We demonstrated the effectiveness of our approach through two case studies conducted with a real-world dataset and positive feedback from domain experts.

Index Terms—Air pollution propagation, pattern mining, graph visualization

1 Introduction

Air pollution has become a global concern due to its severe impacts on many aspects of modern society, such as public health [29] and sustainable development [68]. One of the foremost prerequisites in alleviating air pollution is to understand how pollutants propagate at a large spatial scale, thereby enabling experts to identify the origins and development patterns of the pollution [32, 69]. However, capturing such propagation processes remains highly challenging because of the uncertain pollutant transportation resulting from dynamic wind fields.

With the advancement in data sensing and management technologies [71], data-driven solutions for monitoring, analyzing, and predicting air pollution, such as quality forecasting [66, 74] and local sources discovery [33], are now possible with the large-scale atmospheric data collected by widely distributed weather stations. To analyze the propagation of air pollution, state-of-the-art approach HYSPLIT [52] was put forward in the environmental science literature. This approach was proposed to automatically infer how pollutants are dispersed in the atmosphere and affect an area. Although the potential pollution sources are identified for the given area, HYSPLIT considers neither a) the inherent uncertainties in the dynamic propagation patterns nor b) the complex district-level interactions among multiple cities. Moreover, the lack of an interactive tool that presents the propagation processes from the overview to details at a large spatial scale actively prohibits expert

- *Z. Deng, D. Weng, J. Chen, R. Liu, and Y. Wu are with the State Key Lab of CAD & CG, Zhejiang University. Email: {zikun_rain, dweng, jhchen6, lr_pgd, ycwu}@zju.edu.cn. Y. Wu is the corresponding author.*
- *Z. Wang is with Research Center for Air Pollution and Health, Zhejiang University. E-mail: wangzhibin@zju.edu.cn.*
- *J. Bao and Y. Zheng are with JD Intelligent City Research, Beijing, China. Email: baojie@jd.com and msyuzheng@outlook.com.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

users from comprehending the patterns detected in these processes with their domain knowledge due to the sheer volume of potential patterns.

The aforementioned limitations in the automated approach motivate us to adopt a user-centric analytics approach and develop an interactive visualization system to facilitate the understanding of massive uncertain propagation patterns and identification of major pollution sources and pathways. However, developing such a system poses three challenges.

Extraction of propagation patterns. Understanding how pollutants propagate across multiple districts is the key to finding the cause of air pollution. However, the propagation processes of these pollutants can vary considerably in terms of both space and time. The pollutants in an area can be transported from different sets of polluted upstream areas at different times because wind fields are fluctuating every moment. The transportation speeds also change constantly with the wind speeds at different locations. Hence, an efficient model is required to extract frequent propagation patterns from multistep transportation events and capture the uncertainties of the patterns accurately.

Scalability of pattern presentations. To aid the analysis of the propagation patterns with the visual analytics approach, an intuitive uncertainty-driven visual representation based on directed graphs must be established for these patterns to better illustrate the propagation of air pollution. The problem is that numerous patterns can be extracted from the massive atmospheric data collected at large spatial and temporal scales, demanding considerable effort from users in searching for common and meaningful patterns with traditional graph visualization techniques. Organizing these patterns in a scalable and clutter-free way, which then enables users to identify latent patterns easily with such representations, remains an important yet challenging task.

Analysis of the propagation processes. The propagation processes are massive, spatiotemporal, and uncertain. To assist users in externalizing the pattern identification and analysis procedures, the proposed interactive system must support the effective drill-down exploration [51] in the pattern-instance hierarchy, associating the abstract patterns with physical instances (i.e., pollution events). Moreover, such a system should also incorporate the topology analysis of the propagation processes. For example, star-like propagation structures may indicate that the areas at the center are the major pollution sources. Hence, a novel visualization approach is demanded to facilitate the sophisticated analyses of these processes and reveal in-depth insights.

To address these challenges, we first derive a novel pattern mining framework based on frequent subgraph mining (FSM) method [26] by combining air quality data with meteorological data to model the transportation of air pollutants and detect latent propagation patterns at large spatial and temporal scales. FSM is particularly effective in extracting frequent subgraphs that represent typical propagation patterns. On the basis of this framework, we further propose AirVis, a visual analytics system that enables scalable and intuitive analyses of massive propagation patterns. In particular, we extract significant motifs (i.e., contextless frequent topological structures) from these patterns and aggregate the patterns based on topology similarities with the minimum description length (MDL) principle [46] which compresses data by creating summary representations for similar data items. Moreover, the patterns are organized hierarchically with multiple coordinated views and presented with uncertainty visualization techniques to facilitate the effective exploration and analyses of the pollution propagation processes. To the best of our knowledge, AirVis is the first attempt to establish an interactive topology-driven analysis of uncertain air pollution propagation. Our contributions are summarized as follows.

- ◇ We characterized the problem of analyzing air pollution propagation and compiled a set of analytical requirements based on the iterative discussions with domain experts.
- ◇ We developed a novel FSM-based pattern mining framework that efficiently models the transportation of air pollutants and extracts latent propagation patterns from large-scale atmospheric data.
- ◇ We proposed AirVis, a visual analytics system that enables experts to identify and analyze the patterns in the pollution propagation hierarchically based on the topology with uncertain graph visualizations.

2 Related Work

This section presents prior studies categorized by three relevant topics, namely, model- and data-driven analysis of air pollution, spatiotemporal visualization, and subgraph mining and visualization.

2.1 Model- and Data-Driven Analysis of Air Pollution

Model-driven. Air pollution has been extensively studied in the atmospheric environment literature to analyze and mitigate the pollution [14]. CMAQ [10] is one of the most comprehensive atmospheric dispersion models for analyzing regional air pollution. With the given air pollutant emission and meteorological data, CMAQ computes the concentration of air pollutants based on advection, diffusion, and chemical reactions. For the propagation of air pollution, HYSPLIT [52] is widely used to identify regional pollution sources [30] and propagation pathways [38]. Based on the meteorological data, HYSPLIT attempts to trace back the trajectories of many air parcels starting from a given area (i.e., receptor) for each timestamp. Each of these trajectories is assigned with an estimated concentration value. These trajectories can be further clustered [61] to identify potential patterns or produce a heatmap as a static picture, but no temporal information is provided.

Data-driven. Recent developments in smart city technologies [71] have provided unprecedented opportunities for the data-driven analysis of air pollution with large-scale heterogeneous urban data. Various types of data, including text, image, and traffic data, were used to obtain [34, 39, 73] or predict [16, 66, 74] the air quality. In order to identify pollution sources and propagation patterns, Granger causality [19] has been recently exploited to infer the relationships between different observations [27]. Li et al. [33] constructed massive causality graphs and identified the sources and propagation patterns from them. pg-Causality [76] combined pattern mining techniques and Bayesian learning to capture the causal pathways between neighboring cities.

Although the existing approaches have been proven useful for air pollution analysis, they have limitations. Specifically, the model-driven methods cannot capture the complex district-level interactions (e.g., the pollutants from district A affect both districts B and C), while the data-driven methods fail to consider the continuous propagation across multiple districts (e.g., the pollutants from district A are transported to district D via districts B and C sequentially). The uncertainties of propagation processes are neglected in both types of methods. Moreover, few studies are combined with interactive visualization, leading to obstacles in understanding the propagation. To address the limitations, we propose a novel pattern mining method and design an interactive visualization system to facilitate analyses of air pollution propagation.

2.2 Spatiotemporal Visualization

Spatiotemporal data visualization has been applied in the analysis of air pollution [45, 54, 72], yet no visualization tool is available to analyze pollution propagation. In essence, air pollution propagation can be regarded as a type of movement data. Spatiotemporal visualization for movement data can generally be classified into three categories [1], namely, direct depiction, summarization, and pattern extraction. Specifically, movement data can be directly depicted using visual elements, such as points [17] for OD pairs, polylines [2], tubes [35], stacked bands [55], and space-time cubes [3] for trajectories. Movement data can also be summarized as density maps [36, 47], graphs [58], flow maps [8, 22], OD matrices [21, 65], and OD maps [21]. In addition, pattern extraction can be applied to obtain significant latent patterns from the movement data [12, 24, 62, 75]. However, the existing methods cannot be used to visualize the propagation of air pollution because of its sheer volume and uncertain nature, such as presenting the aggregated uncertainties of the pollutants propagated across districts A, B, and C. To this end, we apply pattern extraction and uncertain graph visualization techniques together with overview-to-detail mechanisms, empowering users to analyze the propagation processes of air pollution interactively.

2.3 Subgraph Mining and Visualization

Frequent subgraph mining (FSM), a classic research topic in data mining [26], has a wide range of applications in various domains (e.g., biology [70], communication [6], and chemistry [15]). Based on the input dataset, existing FSM methods can be categorized into *single graph-based* and *graph transaction-based* methods. The *single graph-based* FSM extracts the patterns that frequently occur within a single but very large graph [57], while the *graph transaction-based*

focuses on those that frequently occur among numerous different graphs named transactions [64]. Our problem is the latter one where the transactions correspond to the massive propagation processes. However, the existing methods like [25] cannot be directly applied to meet our domain-specific problem of extracting significant propagation patterns of air pollution. Thus, we modify the formulation to take into account two aspects of the data, spatiotemporal context and air pollution.

A subgraph is essentially a type of graph. Graph visualization is a broad research field [23, 59]. We focus on the most related parts of visualizing a large number of graphs. Some studies focus on either the structure- [31, 67] or metric-oriented [18, 28] analyses of a large number of graphs, but they cannot be applied to propagation patterns wherein both uncertainties and structure are meaningful. Techniques of dynamic graph visualization [4] aim to organize numerous chronologically evolving graphs (e.g., in a radial [20] or list layout [9, 56]) for efficient time-oriented analysis. These techniques cannot deal with propagation patterns that are not chronologically ordered. Therefore, we propose an MDL-based hierarchical organization of massive patterns exploiting their topology similarities, wherein the uncertainties of the patterns are also visually encoded.

3 Overview

This section introduces the background of our study, describes the relevant concepts and datasets, and summarizes the requirements for the analysis of air pollution propagation.

3.1 Background and Concepts

The rapidly progressing problem of air pollution has become a major issue for many large cities. Among all factors involved in the development of air pollution, the regional transportation of pollutants is generally considered the foremost one, as demonstrated by case studies where environmental scientists analyzed the air pollution problem in Beijing, China [32, 69]. To address this problem, it is important to trace back to the source areas where the pollutants are generated along the propagation pathways and enforce specific policies to mitigate the pollution, such as limiting road traffic and suspending chemical factories. To this end, this paper presents the following concepts to characterize the pattern-based analysis of air pollution propagation and help urban and environmental science experts better understand the transmission of pollutants among multiple cities at a large spatial scale.

- *Transportation* of air pollutants (Fig. 2A). Air pollutants, such as SO_2, $PM_{2.5}$, and PM_{10}, tend to be transported from one district to another via winds. Each process is called an instance of air pollutant transportation or *transportation instance* for short. A transportation instance is described by a tuple with five elements, including the origin and destination locations, the start and end time, and the amount of transported pollutants.
- *Propagation* of air pollution (Fig. 2B). Air pollutants can be transported consecutively across multiple districts. These continuous transportation instances constitute an instance of air pollution propagation or *propagation instance* for short. Numerous propagation instances can be obtained from the atmospheric data owing to the sheer volume of transportation instances.
- *Propagation patterns* of air pollution (Fig. 2C). Propagation patterns are the frequently occurring propagation instances within a spatiotemporal range. For example, in Fig. 2C, a propagation pattern (blue shaded) comprises the most frequent pathways extracted from four different propagation instances distinguished by their colors. Compared with the individual propagation instance that may be a random and unrepresentative event, these patterns strongly indicate the latent pathways of pollution propagation.

3.2 Data Description

Public meteorological and air quality datasets [73, 74] are used in this study. Each dataset includes a *station table* and a *data table*. In the station table, each row represents a meteorological or air quality station identified by its name and district ID (for meteorological stations) or GPS coordinates (for air quality stations). In the data table, each row represents a sample collected in a meteorological or air quality station at a specific timestamp. A meteorological sample comprises weather information, including wind speed and direction, whereas an air quality sample records the concentrations of different air pollutants (e.g., $PM_{2.5}$, SO_2). These samples are collected on an hourly basis.

Without loss of generality, this study focuses mainly on analyzing the propagation of $PM_{2.5}$, a major type of air pollutant that can be transported over long distances [60] and seriously affects public health [63]. After data cleaning and preprocessing, a meteorological sample and an air quality sample for every hour (5,448 hours) were obtained between Sept. 16, 2014 and Apr. 30, 2015 from 138 meteorological stations and 38 air quality stations. We refer to these stations as *districts* because they represent the regional atmospheric conditions.

3.3 Requirement Analysis

To identify the user requirements for analyzing air pollution propagation, we closely worked with three domain experts, EA, EB, and EC, from the fields of urban computing and atmospheric science in the past year. EA and EB have decades of experience in utilizing data-driven automated approaches to study various urban problems, including air pollution, while EC is a recognized environmental scientist specializing in the analysis of air pollution and its propagation. This cooperation attempts to integrate their expertise in evaluating significant propagation patterns and finding major sources of air pollution using the visual analytics approach. To achieve this goal, we closely followed the *nine-stage design study methodology framework* [49]. In particular, we first conducted comprehensive literature review (*learn*) regarding the analysis of pollution propagation and the visualization of spatiotemporal graph data, and then attempted to identify user requirements (*discover*) by holding bi-weekly discussions with the experts for three months. During the development of our system, we iteratively prototyped and implemented several design alternatives (*design & implement*) and evaluated the proposed designs with the experts (*deploy*) to verify the effectiveness of our approach. The derived user requirements are summarized as follows.

R1: Explore a topology-based overview of patterns. To interpret the massive propagation patterns extracted from atmospheric data, the experts must first grasp an overview of these patterns through their topologies, which are the key factors in understanding their roles. For example, a star-like topology indicates that the district at the center of the topology could be a major source of air pollution, while a linked topology represents the long-range transportation of pollutants. Such topology analysis allows the experts to intuitively classify patterns and find promising ones on the basis of their spatial structures.

R2: Obtain the spatiotemporal summaries of patterns. To identify interesting patterns from those with identical topologies, the experts should be able to obtain a clear summary of each propagation pattern in terms of their spatial (*On which path is the pollution propagated?*) and temporal (*When did the associated pollution instances occur?*) dimensions. The summary enables experts to integrate domain knowledge to initially evaluate a propagation pattern, like determining *how a city suffered from the pollution propagated from an industrial city* and *whether it was an abnormity against the meteorological common sense.*

R3: Unfold the uncertain propagation process of a pattern. The experts aim to establish how the pollutants are transported between districts in a specific pattern. In addition to the general overview of propagation pathways, they are particularly interested in analyzing the probabilistic cause-effect relationship between source (*A*) and destination (*B*) districts in the pattern. *How much is the expected amount of pollutant transportation from A to B? Which district is the largest contributor to the pollution in B? Which district has received the largest impact of the pollution in A?*

R4: Find the similarities or differences between patterns. To identify similar or abnormal propagation processes, experts need to compare between these patterns based on the topologies and properties of the patterns. For example, patterns may share the same pollution source but propagate via different pathways. To study the effect of a pollution source, experts need to conduct the comparative analysis based on the cause-effect relationships, including propagation paths and strength, to determine the major pathways. The discovery of similar patterns also allows the experts to analyze them in batch to speed up the workflow.

R5: Examine the propagation instances in a pattern. Each pattern is associated with numerous propagation instances. To obtain reliable and convincing results, the experts need to check these instances for detailed

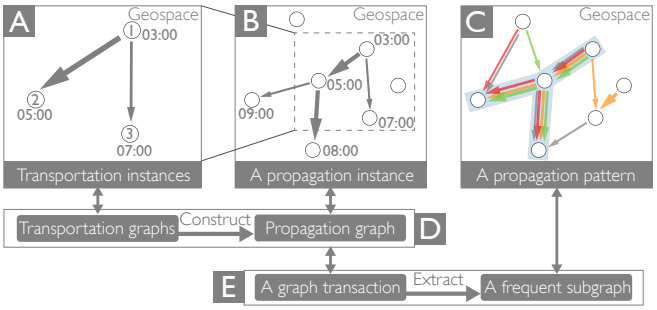

Fig. 2. The pipeline of the proposed frequent pattern mining framework: (A) two transportation instances that describe the pollutants transported between 1 and 2 (larger) and between 1 and 3 (smaller); (B) a propagation instance that involves four transportation instances; (C) a propagation pattern identified from four propagation instances, each encoded with a color; (D) a propagation graph (i.e., a propagation instance) constructed from massive transportation graphs; (E) a frequent subgraph (i.e., a propagation pattern) extracted from massive graph transactions.

information, e.g., how frequently did the relevant pollution events occur and did the pollution become severe over time? Furthermore, displaying the raw samples with numeric readings is desirable because the experts are familiar with such a presentation.

3.4 System Architecture

Based on the aforementioned requirements, we develop AirVis which comprise two parts, namely, the backend and the frontend. The backend of the system, implemented in Node.js, integrates an FSM-based pattern mining model to extract frequent propagation patterns efficiently based on the parameters specified by the users. The frontend, written in Vue.js and TypeScript, runs in modern web browsers and allows users to interactively explore and analyze the complex propagation patterns, which are revealed with the hierarchical visualizations based on the topologies, propagation processes, and instances of the patterns.

4 Pattern Mining

This section describes a novel mining framework for the efficient extraction of air pollution propagation patterns. First, we model the transportation of air pollutants quantitatively. Then, we construct propagation graphs based on the modelled transportation instances. Finally, we leverage the FSM method to extract the propagation patterns.

4.1 Modelling Pollutant Transportation

The transportation of air pollutants between two districts is highly uncertain due to the varying wind fields. To model such a process from district i to j, for example, we need to determine the probability of the air pollutants transported from i to j (Fig. 3A). By leveraging the *air parcel* concept from the environmental science literature [50], we attempt to simulate the movement of air pollutants with air parcels based on a quantitative sampling method (Fig. 3C). Specifically, s air parcels, representing the air pollutants, are released near the district i at the timestamp t in a simulation. The locations of these air parcels are updated iteratively based on the meteorological conditions, including the wind speed and direction (Fig. 3B), until the distance between the air parcel and the district j falls under $d_e = 20\,\text{km}$ or the time limit is exceeded. Inspired by HYSPLIT [52], we update the location of an air parcel with the following equations at the timestamp t:

$$L(\Delta t + t) = L(t) + \vec{v}_t \times \Delta t$$
$$\vec{v}_t = \frac{\sum_{m \in M}(d_n - d_m) \times \vec{v_{m,t}}}{\sum_{m \in M}(d_n - d_m)},$$

where $\vec{v}_t$ is the velocity of an air parcel at the location $L(t)$ at the timestamp t, and M is a set of neighboring meteorological stations, where the distance between each station and the air parcel is measured to be d_m, within the given distance threshold $d_n = 30\,\text{km}$. The nearer the station is, the more its observed wind vector $\vec{v_{m,t}}$ contributes. Thereafter, we denote the number of the air parcels that reached the district j as s_r and the travel time of the k-th air parcels as tt_k, $k \leq s_r$. Hence, we

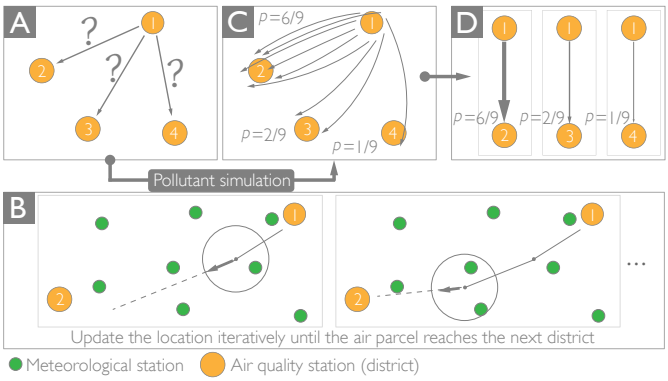

Fig. 3. (A) The probabilities of the air pollutants transported from 1. (B) The iterative update of the air parcel locations. (C) The quantitative sampling method. (D) The extracted transportation instances.

estimate the transportation probability p_{ijt} from district i to j at the timestamp t with the empirical observations:

$$p_{ijt} = \frac{s_r}{s}.$$

The average transportation time tt_{ijt} is calculated to be $(\sum_{k=1}^{s_r} tt_k)/s_r$. Fig. 3C and 3D demonstrate an example of inferring transportation probabilities from air parcel sampling.

Subsequently, we represent each transportation instance with a *transportation graph* (Fig. 2D). Each transportation graph has two nodes n_i and n_j that represent two respective districts i and j and one directed edge e_{ij} indicating the transportation process. In addition to the inherent attributes, such as the name and geo-location, each node n has two extra attributes, namely, the timestamp $n.t$ and the pollution concentration $n.c$ observed at the timestamp $n.t$. Furthermore, we compute the following attributes and associate them with the edge e_{ij}: a) the transportation time of the transportation instance is denoted with $e_{ij}.tt$; b) the estimated transportation probability is denoted with $e_{ij}.p$, which is also the *contribution* factor of this transportation instance that describes the ratio of the pollutants transported from the source to the destination district; c) the expected amount of the transported pollutants is derived with $e_{ij}.tc = e_{ij}.p \times n_i.c$; and d) the *impact* factor $e_{ij}.a = e_{ij}.tc/n_j.c$ indicates the ratio of the pollutants in the destination that are received from the source district. Based on experts' suggestions, we removed the graphs with $e_{ij}.tc < 30$, assuming that no pollutant is transported.

4.2 Constructing Propagation Graphs

The transportation instances only describe the pollutant transmission from one district to another. To characterize the pollutants propagated among multiple districts, we derive the propagation instances, represented by *propagation graphs* (Fig. 2D), by searching and merging the extracted transportation graphs.

To better illustrate the construction of propagation graphs, we first define the concept of *spatiotemporal continuity* (Fig. 4): two transportation graphs are spatiotemporally continuous iff they share one and only one node, where the locations and timestamps of the shared nodes are equal, respectively. Based on this concept, we follow a depth-first search procedure outlined as follows: starting from a random transportation graph, we recursively add the transportation graphs that are spatiotemporally continuous with the added ones to the propagation graph. To simulate the decomposition of air pollutants, we limit the maximum time span of the propagation graph to 200 hours, as recommended by the experts. By repeating this procedure, we are able to generate a massive number of propagation graphs $pG = (V, E)$ until all the transportation graphs have been visited. The attributes associated with the nodes and edges in the propagation graphs are identical to those in the transportation graphs. Such method allows us to obtain the continuous long-distance propagation instances naturally as paths in the graphs, which are almost impossible to detect with the prior methods.

4.3 Extracting Propagation Patterns

To extract the meaningful patterns from the propagation instances, we seek to find the persistent and significant parts in the generated propaga-

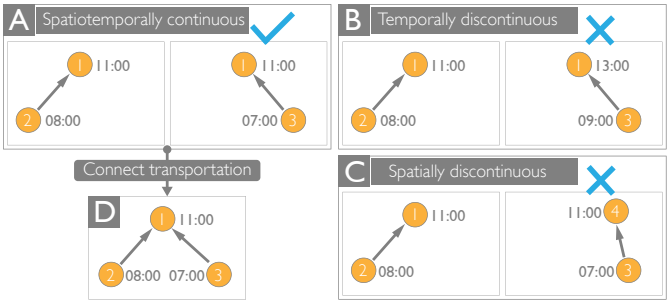

Fig. 4. (A) Spatiotemporally continuous transportation instances are connected as a propagation graph in (D). (B) and (C) illustrate the temporally and spatially discontinuous transportation instances, respectively.

tion graphs with the frequent subgraph mining (FSM) techniques. We convert the extraction of the propagation patterns to an FSM problem with the following concepts (Fig. 2E):

- *Transaction*: Transactions are the input of the FSM problem. Each transaction is defined as a graph that comprises a set of labeled nodes and edges. In our scenario, we refer to the propagation graphs as the transactions.
- *Support and frequent subgraph*: A subgraph g (i.e., a propagation pattern) may match a number of transactions based on the graph isomorphism. The *support* of g is the number of the matching transactions divided by the number of total transactions. g is frequent iff its support is larger than a given threshold λ_c.

We employ a well-established Apriori-based method [25] to extract the frequent subgraphs. In addition, we allow users to interactively control the pattern mining process by setting a transported pollutant threshold λ_p to exclude from support computation the propagation graphs where the pollutants transported on the matched edges $e.c$ are below the threshold. Furthermore, the support is calculated with the time span covered by the matched transactions divided by the total time span to emphasize the importance of the pollution durations.

Finally, the massive propagation patterns (i.e., frequent subgraphs) can be efficiently extracted. Each pattern $P = (V, E)$ comprises a node set V comprising the involved districts and an edge set E comprising the propagation pathways. Each node $v_i \in V$ maintains a list of the pollution concentrations $v_i.C$ collected from the corresponding nodes in the matched propagation graphs. Similarly, each pathway $w_{ij} \in E$ from district i to j is associated with $w_{ij}.TT$, $w_{ij}.P$, $w_{ij}.TC$, $w_{ij}.A$, lists of transportation times, transportation probabilities, transported pollutant concentrations, impact factors, respectively. Please refer to the appendices for all used notations and details of pattern extraction.

5 Visual Design

In this section, we present the design goals compiled from the aforementioned user requirements and describe the design of AirVis, a visual analytics system for analyzing the propagation of air pollution.

5.1 Design Goals

To satisfy the user requirements and support the analytical tasks, we further compile a set of design goals summarized as follows:

G1: Hierarchical presentation of propagation patterns. To address scalability issues and facilitate the effective exploration of massive patterns (R1, R2, R3, and R5), we follow the *visual information-seeking mantra* [51] and organize these patterns hierarchically with multiple coordinated views. Users should be able to identify interesting topologies first (R1), browse through the patterns with an identical topology (R2), proceed to the analysis of a particular pattern (R3), and finally dive into the details of the associated propagation instances (R5). Moreover, the design should also support the intuitive filtering of patterns.

G2: Uncertainty-aware visualization of pattern topologies. Patterns with similar propagation topologies should be aggregated to provide a comprehensive overview (R1). In particular, each edge in the aggregated topologies comprises a set of propagation probabilities associated with the corresponding edges of the pattern. These probabilities should be reflected on the topology visualizations to assist users in determining the stability of propagation structures.

G3: Spatiotemporal visual summaries of patterns. To help users intuitively learn the basic characteristics of the patterns, the system should compactly present the visual summaries of these patterns in terms of their spatial contexts and temporal distribution (R2). Glyph is a ideal candidate since many visualization studies [5] have demonstrated its effectiveness in multifaceted analysis, where complex data can be encoded with different visual channels in a space-efficient way.

G4: Intuitive illustrations of propagation processes. The propagation processes of a pattern (R2) are based on directed graphs, where the nodes represent the involved districts and the edges indicate the transportation of pollutants. Thus, these processes can be intuitively visualized with a node-link diagram on the map. Moreover, the probabilistic cause-effect relationship between the districts can be encoded on the node-link diagram to help users interpret the pattern of interest.

G5: Comparative visual analysis of propagation patterns. The proposed design should allow users to compare among multiple interesting patterns according to their topologies and attributes (e.g., expected amount of propagated pollution) and find their similarities and differences (R4). This comparison can likewise be facilitated further by automated approaches, such as dimensionality reduction techniques, to enable the effective discovery of pattern clusters and outliers.

G6: Detailed inspection of relational propagation instances. To support the analysis of the propagation instances associated with a pattern (R5), the proposed design should a) enable users to select a time frame of interest on the basis of the temporal distribution of the instances and b) present the propagation relationships among involved districts in a scalable fashion. Moreover, the raw readings of air quality data should be included in the proposed design.

Based on these design goals, we develop AirVis to assist the expert users in visually analyzing and sensemaking the uncertain propagation of air pollution at a large spatial scale. AirVis follows a hierarchical exploration scheme (G1) to facilitate the effective visual analysis:

Topology Visualization (G2): The topologies of the extracted propagation patterns are aggregated and visualized in the *motif view* (Fig. 1B) based on the MDL principle and the idea of motifs [42]. Moreover, the uncertainties of the topology structures are encoded within the glyphs.

Pattern Visualization (G3, G4, and G5): By expanding a motif, users can browse through the associated pattern glyphs in the *pattern view* (Fig. 1C) and learn the complex propagation processes of the selected patterns via the *inspection list* (Fig. 1C3), where the juxtaposed propagation graphs provide the support for intuitive side-by-side comparative analysis. In addition, an automatically generated *projection view* (Fig. 1A3) is incorporated to help users identify similar patterns.

Instance Visualization (G6): The instances in the selected pattern are depicted in the *instance view* (Fig. 1D). By selecting a timeframe according to the temporal distribution of instances, users can explore the transportation of pollutants among the involved districts with the chord diagram and inspect the detailed numeric data in the table.

5.2 Topology Visualization

To facilitate the topology-driven analysis of air pollution propagation (G2), we aggregate similar patterns based on topology into motifs and visualize these motifs with uncertainty-aware visual encodings.

5.2.1 MDL-Based Pattern Aggregation

Inspired by the concept of *motifs* in the field of network analysis [42], we denote the contextless topological substructures in the propagation graph of a pattern as motifs. To detect patterns with the similar topologies, *significant motifs* (Fig. 5) that are commonly shared among the patterns can be extracted as the representations of the corresponding patterns, enabling users to obtain the key topological characteristics of the underlying patterns intuitively. Therefore, such motif extraction should satisfy the following two requirements: a) *generality*: the significant motifs should be general to represent a large portion of the pattern topologies; and b) *similarity*: each significant motif and its corresponding patterns should be identical to avoid undesired aggregation.

We extract these motifs based on the minimum description length (MDL) principle [46], which has been demonstrated to be particularly effective in sequence summarization [13] and graph simplification [11]. The MDL principle seeks to find a common *summary* S_i for a group of data items and denote the remaining part of the j-th item in the

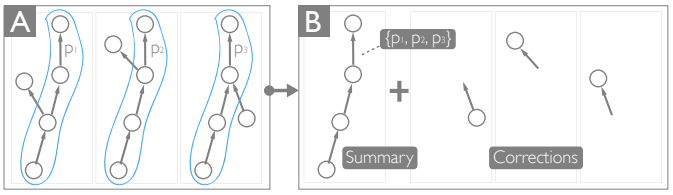

Fig. 5. The extraction of motifs: (A) three structurally similar propagation graphs, where the summary topology is circled in blue; (B) the summary (i.e., significant motif) and corrections generated by the MDL principle.

group as the correction C_{ij}. By minimizing the overall description length $L = \sum_i(|S_i| + \sum_j |C_{ij}|)$ for all groups, we can obtain compact data representations that satisfy the aforementioned requirements.

To adapt the MDL principle to the extraction of the motifs, we first deduplicate the propagation graphs to remove the graphs with the duplicate topologies and then use the remaining graphs as the input data items $G_i = (V_i, E_i) \in G$ to find a set of summaries $S = \{S_1, S_2, ...\}$, where S_i is a motif that represents a series of propagation graphs $\{G_{s_{i1}}, G_{s_{i2}}, ...\}$ of length k_i, i.e., $S_i \subseteq G_{s_{ij}}$. Similarly, the correction C_{ij} is defined as $G_{s_{ij}} \setminus S_i$, as illustrated in Fig. 5B. Based on the generality and similarity requirements, our goal is to minimize:

$$L(G,S) = \sum_i L_i(G,S_i) = \sum_i (|S_i| + \sum_j |C_{ij}|),$$

where $|S_i|$ and $|C_{ij}|$ denote the number of edges in the respective graphs. The greedy algorithm proposed by Navlakha et al. [43] can be applied to calculating S by iteratively merging the pair of the motifs S_i and S_j that maximizes the decrease of $L(G)$, until no such pair exists. In addition, we introduce a constraint $L_i(G,S_i)/k_i > L_m(G,S_m)/k_m$ and $L_i(G,S_j)/k_j > L_m(G,S_m)/k_m$ based on the average of description lengths, where S_m is the summary after merging and $k_m = k_i + k_j$ is the number of the patterns associated with the motif S_m. This constraint allows the resulting motifs to bias towards the similarity requirement, thereby generating more accurate topological representations. Please refer to the appendix C for more details.

The computation of the summary set S can be further accelerated with a tree index, where we start from the root corresponding to a graph with only one vertex and generate each child node by iteratively adding an edge to the graph associated with the node's parent until the index contains all propagation graphs. Subsequently, we remove the nodes that are not a part of any propagation graph. With such index, the summary set can be efficiently approximated by selecting the node that contains the most propagation graphs and merging all of its propagation graphs greedily in batch for each level of the tree from the bottom up.

5.2.2 Uncertainty-Aware Motif Visualization

The extracted significant motifs are then visualized in the motif view (Fig. 1B) as the summary representations of pattern topologies. To help users identify the structures of the motifs intuitively, we present these motifs with node-link diagrams based on the force-directed layout algorithm (e.g., Fig. 1B4), where the corrections of the associated propagation graphs are superimposed on each extracted motif with the relatively smaller node size and lighter color.

Uncertainty-aware glyphs. Each edge of the significant motifs comprises a series of propagation probabilities corresponding to each edge of the associated propagation graphs. Hence, we visualize these probabilities on each edge of the motifs with an intuitive fusiform glyph (Fig. 1B3) indicating the mean and variance of the probability distribution. The mean of the probabilities, encoded with the opacity of the glyph, reveals how likely the pollutants will be propagated along the edge, while the variance, encoded with the width of the glyph, helps users determine the stability of the edge in such propagation structure. Such glyph design allows users to quickly identify the strong and persistent propagation structures from a list of motifs. In addition, the motifs can be sorted by the average mean or variance of their edges.

Design alternatives. An alternative to the aforementioned glyph design is to encode the probability distribution directly along the edges with the summarization representations like heatmaps (Fig. 6A). Users can obtain the fine-grained propagation probability information from the motifs. However, two limitations are observed in such design: a) the conflicting directions of the edges introduce the difficulty in the readability of the glyphs, i.e., the starting points of the axis, and b) users complained that they had misinterpreted this visual encoding as the spatial distribution of pollutants among the involved districts. Hence, we have simplified the design of the glyphs and selected two representative features, the mean and the variance, to help users intuitively grasp the uncertainty in the topological structures of the significant motifs.

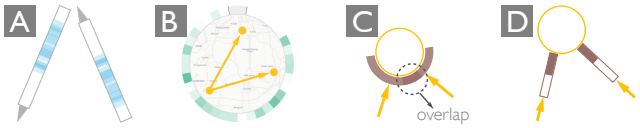

Fig. 6. The design alternatives to (A) the uncertainty visualiztion in the motif glyphs, (B) pattern glyphs, and (C) (D) pattern graphs.

5.3 Pattern Visualization

By selecting a significant motif in the motif view, users can analyze the propagation patterns associated with the selected motif in the pattern view (Fig. 1C). However, presenting all patterns in the same spatial context will result in overlapping edges and severe visual clutters, and such method cannot provide important insights into multi-district propagation processes. Moreover, these patterns are closely related to their spatiotemporal context and associated with multiple probabilistic attributes including the contribution and impact factors, and thus the existing graph visualization methods [44, 48, 58] cannot be directly applied. To help users locate interesting patterns and inspect the detailed propagation processes, we adopt a visual representation with two levels of detail for the patterns. The first level of detail (Fig. 1C1) comprises a list of compact *pattern glyphs* that outline the spatiotemporal information of each pattern, with which users can obtain a brief overview about the spatial contexts and temporal distributions of the patterns (i.e., when and where the pattern occurred) (G3). Thereafter, users can add a pattern to the inspection list (Fig. 1C4), where the detailed propagation processes of each added pattern will be depicted with a *pattern graph* on the map served as the second level of detail (G4). Additionally, in the projection view, we lay out the patterns according to their similarities computed based on the word2vec model [40, 41] to assist users in identifying pattern clusters and outliers efficiently (G5).

5.3.1 Pattern Glyphs

The design of the pattern glyphs is illustrated in Fig. 1C1. The small dots enclosed in the large circle represent the districts and are laid out according to their geospatial positions. The districts and pathways involved in the propagation process are highlighted in orange, while the irrelevant dots are rendered in gray. The red arrow on the border points towards the average direction of the propagation pathways. To facilitate the visibility of the propagation process in such a compact space, the fisheye effect is applied on mouse hover (Fig. 1C3).

The temporal distribution of the propagation instances is depicted with a bucketed heatmap around the glyph. Each bucket represents a week, resulting in total 33 buckets. The top of the heatmap is explicitly made discontinuous to avoid the confusion that the distribution is cyclic.

Design alternatives. Instead of the dot-based representation, an alternative is to embed a map directly in the glyph focusing on the spatial region that contains the propagation process (Fig. 6B). However, the readability of the map is limited because of the compact space in the glyphs, and the spatial contexts of the glyphs are difficult to compare due to the lack of a consistent spatial reference.

5.3.2 Pattern Graph

The pattern graphs are designed to visualize the uncertain propagation processes of a pattern on the map. Each transportation pathway between two districts (e.g., the source district A and the destination district B) in the propagation processes comprises two key features, the contribution and impact, extracted from all transportation instances. The contribution of a pathway includes all expected ratios of the pollutants in A contributing to B. Similarly, the impact comprises the expected ratios of the pollutants in B received from A.

We illustrate such cause-effect relationships with a tailored node-link diagram (Fig. 1C5). The districts involved in the propagation process

are represented with the circles filled with yellow on the white background. Moreover, the size of these circles indicates the concentration of pollutants. Around the circle, each small pie chart in red or green is linked with an incoming or outgoing edge and encodes the median of the impact or contribution ratios with the filled part, respectively. In particular, the expected ratios are represented with their medians, since medians are less sensitive to extreme values.

Design alternatives. Besides the *satellite* layout proposed above, we developed two alternative layouts, the *circumferential* and *radial* layouts, to encode two percentage values at both ends of each edge in the node-link diagrams. The circumferential layout (Fig. 6C) encodes the values as donut slices along the circumferential direction of the nodes. However, such encoding may result in severe occlusions between slices. The radial layout (Fig. 6D) depicts the values along the radial direction with bars. Nevertheless, such encoding has the scalability issue because each glyph requires more screen space. Hence, we chose the satellite layout to visualize the percentage values compactly and intuitively.

5.3.3 Pattern Similarity Detection

In addition to the juxtaposed comparative analysis in the inspection list, the projection view (Fig. 1A3) is designed to assist users in understanding the similarities and differences among the patterns and identifying the interesting pattern clusters and outliers efficiently.

Vector representation. To compute the similarities, we first obtain a semantic vector representation for each pattern based on the word2vec model. The word2vec model has been demonstrated to be highly effective in generating a vector for each word in the vocabulary extracted from a series of sentences while the Euclidean distance between two vectors indicates the similarity between two corresponding words. In our scenario (Fig. 7C), we use the patterns as the words and the transactions (i.e., the propagation instances that support the patterns) as the sentences, such that we can directly feed the patterns and transactions into the word2vec model and obtain the vectors that represent the patterns. In particular, a unique ID is assigned to each pattern, and the feature vector for each transaction is generated with a list of pattern IDs the transaction belongs to (Fig. 7C1). We set the scanning window to the largest number of patterns in a transaction in order to make the patterns in the same transaction related to each other.

Visualization. Given the vector representations, we perform the dimensionality reduction with t-SNE [37] to preserve the local similarities among the patterns and obtain the two-dimensional coordinates of the patterns. These patterns are subsequently plotted as a scatterplot, where the mean concentration of the estimated transported pollutants in all districts is encoded with the opacity of the points for each pattern.

Alternatives. We also attempted to perform the t-SNE method directly with one-hot vectors (Fig. 7A and 7B). These one-hot vectors are generated either by concatenating the features (e.g., the time occurred, district, and pollution concentrations) or with the transactions directly. However, the results failed to capture the similarities among the patterns accurately, mainly because these vectors are too sparse.

5.4 Instance Visualization

By selecting an interesting pattern, experts can inspect its detailed propagation instances (G6) via the instance view (Fig. 1D).

Pollutant transportation. The transportation instances among the involved districts within the timeframe selected on the slider are visualized with a chord diagram (Fig. 1D1), which offers the occlusion-free exploration of the instances in terms of both spatial and temporal dimensions. Each district is represented with an arc, along which the time-varying pollutant concentrations are encoded in a clockwise direction with the height and luminance of the bars. The transportation instances between pairs of districts are indicated with chords, the widths of which encode the impacts of the instances.

Raw instance data. An *instance table* (Fig. 1D2), coordinated with the chord diagram, is designed to provide the raw data of the propagation instances with numeric values and the distributions in boxplots. The raw data of each district pair in the propagation instances is depicted with a nested table (e.g., the statistics of pollutant transportation from Xinjishi to Shenzexian are shown in Fig. 1D3). These raw data include the pollutant concentrations in the districts, propagation probability, transportation time, and occurred time.

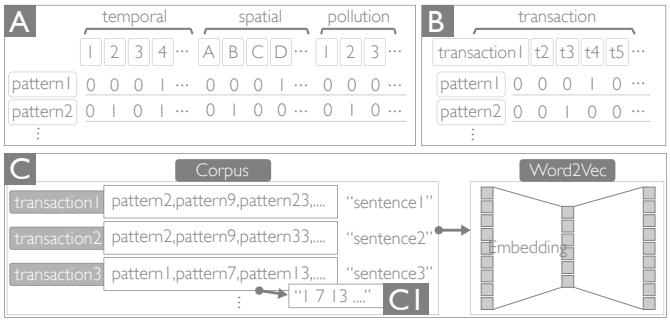

Fig. 7. Three methods to generate the vector representation of a pattern, (A) concatenating the one-hot encoded feature vectors, (B) using transactions with one-hot encodings directly, and (C) learning the vector representations of the patterns based on the word2vec model.

5.5 Interactions

The following three interactions are incorporated into the proposed system to further enhance its usability.

Interactive mining. Our system allows users to interactively tune the parameters in the mining model, including the time span, λ_p, and λ_c, and obtain the generated patterns in real-time with a control panel (Fig. 1A1). These parameters default to less strict values, thereby allowing the model to produce a reasonable number of patterns. Users can customize these parameters according to their needs.

Pattern filtering. Users can select a district and filter out the patterns that does not involve the selected district. Moreover, the selected district will be highlighted in the motif and pattern glyphs. Users can also draw a polygon selection on the projection view to see the desired patterns.

Hierarchical exploration. Users can unfold the motif and pattern glyphs to reveal more details on demand. The interesting patterns can be added to the inspection list, where these patterns can be compared side by side and further inspected in the instance view.

6 Evaluation

We demonstrated the effectiveness and usability of AirVis via two case studies and the interviews with three domain experts (EA, EB, and EC). Before the case studies, we conducted a training session to walk the experts through our system, including the visual encodings and user interactions. Thereafter, the experts explored the propagation patterns with the system and investigated their subjects of interest: a) the propagation patterns in the North China Plain, and b) the propagation patterns that involved Beijing. In these case studies, the experts obtained valuable insights that could potentially alleviate the air pollution problem and guide the pollution control policies in China. We then interviewed the experts to collect their comments and feedback.

6.1 Regional Analyses

The North China Plain is not only the political center but also the most polluted area in China. The experts aimed to identify the districts that acted as the pollution sources in air pollution propagation and understand how the districts interacted with one another.

The experts opened the system and immediately noticed a heavily colored pattern cluster at the top left of the projection view (Fig. 1A3), which represented a group of patterns with high pollutant concentrations. Given that the projection view captures the similarities of the patterns based on their spatiotemporal contexts, this pattern cluster indicates that the propagation of air pollution with high pollutant concentrations occurs consistently in the same spatiotemporal domain. Such an observation triggered the experts' interest. After these patterns were selected by drawing a polygon, a set of significant motifs were displayed in the motif view (Fig. 1B). Most of these motifs comprised four or even five nodes, indicating that this cluster of propagation patterns involved many districts and covered a large area.

Chain-like motif exploration. The experts noticed a motif with a chain-like topology (Fig. 1B1), and each edge of this motif had a nearly-opaque but wide fusiform glyph indicating both of the high transportation probability and uncertainty. Curious about this type of

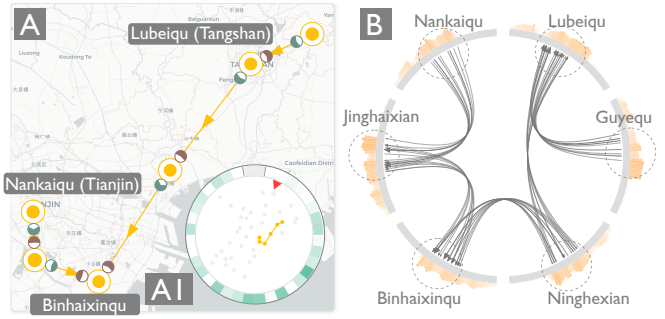

Fig. 8. The propagation pattern involving Binhaixinqu. (A) Air pollutants are propagated from Tangshan and the downtown of Tianjin. (B) Significant pollution correlations among the districts (dashed circles enclose the correlated temporal ranges among these districts).

propagation, the experts clicked on the motif and retrieved a set of the propagation patterns with such topology in the pattern view (Fig. 1C).

The experts skimmed through these patterns and quickly acquired a general impression. In particular, all patterns in the pattern view (Fig. 1C) exhibited an almost linear shape spatially and spread across the southwestern and northeastern regions. The experts suggested that the low uncertainty of such topology might result from the constant and strong winds along these paths. Moreover, the temporal distributions and the propagation direction arrows also showed that the propagation patterns from southwest to northeast (e.g., the first two patterns in Fig. 1C) mainly occurred in autumn and early winter, whereas those in the reversed direction (e.g., the last three patterns in Fig. 1C) often occurred in late winter and spring. This phenomenon is consistent with the meteorological conditions. While browsing the spatial contexts inside the glyphs, the experts found two patterns (i.e., the first two patterns in Fig. 1C) that started near Shijiazhuang in winter and autumn and propagated the air pollutants towards Langfang along two different but almost parallel paths. The experts were interested in these patterns because there were many coal-fired power plants around Shijiazhuang.

Experts further unfolded these patterns to study their propagation processes. In the pattern graphs illustrated in Fig. 1C5 and 1C6, two districts, Xinjishi and Xinleshi, were identified as the origins of air pollution. The large sizes of the inner orange circles suggested that these two districts were seriously polluted. Moreover, the pie charts in Fig. 1C5 and 1C6 indicated that the pollution propagated along two different paths was largely contributed by the pollution in these two districts, which were constantly exposed to the pollution from Shijiazhuang. EA and EB hypothesized that the coal-fired power plants could be a severe air pollution source that had a considerable impact on remote regions. This hypothesis needs to be confirmed with advanced chemical reaction analysis. Moreover, EC also indicated that the pollution in Shijiazhuang might reach northeastern districts, including Beijing and Langfang, via these two paths. Therefore, informed pollution control policies can be made with these patterns, such as suspending a few power plants in Shijiazhuang based on the forecasted wind conditions, which has been proven to be particularly effective [53].

Unbalanced motif exploration. EB noticed another stable motif with an interesting structure (Fig. 1B2) wherein a district was polluted by the pollutants propagated along two different paths, one of which was relatively long. Such structure indicated that this district had a higher vulnerability to air pollution. EB selected this motif to obtain the associated patterns, where he identified an interesting pattern that occurred in two large cities, Tianjin and Tangshan (Fig. 8A1). In the corresponding pattern graph (Fig. 8A), the propagation pathways suggested that the pollution in the Binhaixinqu district was likely propagated from the downtown of Tianjin and a remote city, Tangshan. Fig. 8B shows the deterioration of air quality among the involved districts is highly correlated in the time ranges enclosed with dashed circles, which confirmed the aforementioned hypothesis. However, EB indicated that alleviating the pollution in Binhaixinqu was especially challenging because such pollution involved multiple sources.

Medium-polluted pattern analysis. Furthermore, EA also wanted to examine the patterns with medium air pollution that were represented by the relatively transparent points sparsely distributed in the

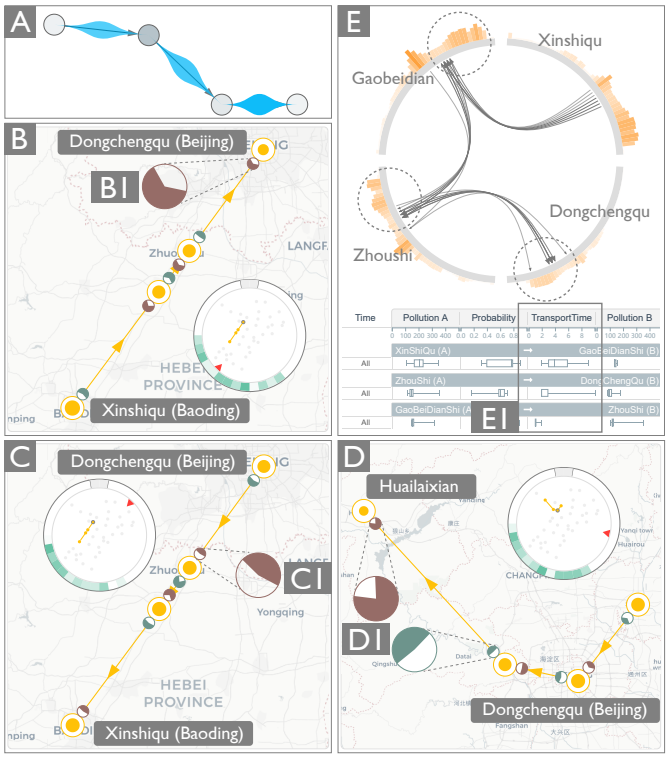

Fig. 9. (A) The significant motif that involves Beijing. There are three types of the propagation processes involving Beijing, namely, those (B) from Baoding to Beijing, (C) from Beijing to Baoding, and (D) from Beijing to Huailaixian. (E) The propagation instances from Baoding to Beijing.

projection view (Fig. 1A4). After selecting these patterns, only one significant motif that comprised two nodes and one edge was shown in the motif view (Fig. 1B5). Compared with the previously inspected patterns of serious air pollution, each of the current patterns involved fewer districts and covered a smaller area. EA suggested that this was because the pollution could propagate to remote districts only if the pollution was severe. Otherwise, the pollutants would be decomposed during the long-distance propagation processes. These patterns also had diverse temporal distributions and were sparsely distributed on the map (Fig. 1C7). This observation is consistent with the scattered points in the projection view, demonstrating that our projection method is successful in capturing the spatiotemporal similarities of the patterns. The experts told us that these patterns were likely to reflect the interactions of local pollution, such as traffic emissions, among different cities.

6.2 District-Centered Analyses

The cause of air pollution in Beijing, the capital of China, still remains controversial [32]. Some researchers concluded that regional transport of $PM_{2.5}$ is the main cause of air pollution in Beijing, while others attribute the air pollution to local production. The experts tried to answer this question by exploring propagation patterns with AirVis.

In particular, they wanted to investigate the significant propagation patterns of air pollution in winter, which is the time when the air pollution in Beijing becomes the most serious. Therefore, the experts set the time span to winter and increased the threshold of support λ_c to 0.05. According to these constraints, a new set of more significant patterns were extracted. The representative air quality station of Beijing is located at Dongchengqu, a center district of Beijing. Hence, the experts filtered out the irrelevant patterns that do not involve Dongchengqu (Fig. 1A2). Given that a motif represents multiple topologically identical patterns, the roles of the Dongchengqu in the propagation processes are encoded with the grayness of the corresponding nodes based on the number of topological matches. The experts then identified a group of patterns in which Dongchengqu acted as both of the pollution source affecting several downstream districts and the victim exposed to the air pollutants from upstream districts (Fig. 9A). After comparing this group of patterns, the experts disclosed three main

types of propagation processes:

1) **From Baoding to Bejing** (Fig. 9B). There are some patterns starting from Baoding and ending in the downtown of Beijing. The experts learned that the external pollution in Beijing is likely from Baoding, so they carefully inspected this type of patterns in the instance view (Fig. 9E) to check the detailed effects and transportation times (Fig. 9E1).They found that such patterns would result in severe air pollution along the propagation pathway, as shown in the temporal distribution of air quality (enclosed in dashed circles in Fig. 9E). By accumulating the average of three transportation time boxplots in Fig. 9E1, the experts inferred that the pollution in Baoding will be propagated to Beijing in approximately 7 hours. This information is critical for the local authorities to issue early air quality warnings.

2) **From Beijing to Baoding** (Fig. 9C). Surprisingly, Beijing also served as a source that polluted Baoding. In contrast to the propagation processes described above (Fig. 9B), the pollution in Beijing were propagated along the exactly same path but in the opposite direction. Moreover, the experts noted that the impact factors in these processes (Fig. 9C1) were smaller than those shown in Fig. 9B1, indicating that the pollution from Beijing were not the main cause of the pollution in the downstream districts. EB also told us that most of the pollution locally produced in Beijing came from traffic emissions, which were less serious than the industrial emissions in southern districts.

3) **From Beijing to Huailaixian** (Fig. 9D). In addition, the pollution in Beijing also affected Huailaixian in the northwest region. According to EA, Huailaixian is a clean city without heavy local emissions. The pie charts in the pattern graphs (Fig. 9D1) implied that less than the half of the pollutants from Beijing had caused a large portion of the pollution in Huailaixian. The experts also investigated this pattern extensively in the instance view and confirmed the hypothesis.

In conclusion, air pollution in Beijing is from both internal and external sources while the external sources are more influential. Such finding is consistent with the environmental science literature [32].

6.3 Expert Interviews

After the case studies, we conducted informal interviews with the experts and gathered their comments and feedback regarding the visual design, user interactions, and usability of the proposed system.

Visual design. All three experts agreed that AirVis could intuitively depict the complex propagation processes of air pollution with the motif glyphs, pattern glyphs, and pattern graphs, and present the detailed information of propagation instances in terms of both spatial and temporal dimensions. "*Many interesting patterns are revealed with the topology-pattern-instance hierarchy.*" commented EA, "*I believe our current studies will benefit from this system.*" EA and EB also praised our system for being "*aesthetically appealing*".

User interactions. All three experts indicated that the interactions implemented in our system were very smooth. EC was particularly impressed by the interactive pattern mining, where he could specify a $PM_{2.5}$ threshold and the system would return the extracted propagation patterns in real-time. Moreover, EA and EB found that the district filter was highly convenient in the district-centered analysis.

Usability. All three experts confirmed the usability of our system. "*The spatial granularity is much finer than the existing approaches, which makes it possible to obtain deeper insights with this system.*" commented EC. Furthermore, EC suggested that we could further improve the transportation simulation by adopting an ensemble model that incorporates the results generated by HYSPLIT [52].

7 Discussion

This section discusses the implications and limitations of our work. We also share several important design lessons we learned.

Implications. To the best of our knowledge, AirVis is the first visual analytics system that efficiently incorporates users' expertise in the domain of air pollution propagation analysis. The availability of AirVis significantly promotes the efficient exploration of large-scale pollution propagation data and the acquisition of deep insights into the uncertain propagation patterns. Moreover, our work is largely distinguished from the state-of-the-art approaches by improving the spatial granularity of air pollution propagation analysis to the station level and the temporal granularity to an hourly resolution, and the topology detection and uncertainty analysis features provided by our system enable experts to identify and analyze new propagation patterns more effectively compared with the existing tools. Such fine-grained analysis allows our system to precisely capture the continuous propagation of air pollution and facilitate the development of timely and accurate pollution control policies, which may help reduce economic losses and even save lives.

Our approach can be generalized to address the similar problems that involve the topology-driven analysis of massive small graphs. Applying the MDL principle in extracting motifs effectively reduces the number of graphs and summarizes these graphs by their corresponding topological representations with the corrections as visual hints. In addition, the hierarchical exploration scheme we followed while designing our system largely allievates users' cognitive load in analyzing numerous graphs and enables users to progressively obtain the overview of the patterns and analyze the interesting details on demand.

Limitations. Three limitations are observed in our work. First, a vectorized wind field is reconstructed from sparse meteorological samples to simulate the air pollutant transportation. The sparsity may result in noticeable inaccuracies in the simulation, particularly in the areas with complex geographical conditions. This limitation can be addressed by incorporating the fine-grained wind field data or a more sophisticated interpolation method based on the terrain data. Second, the distance thresholds used in the pollutant transportation modelling is determined in advance based on the experts' knowledge and the density of monitoring stations. To establish more accurate modelling, an interactive interface can be incorporated in the future work to allow users to fine-tune the thresholds. Third, our approach simply estimates the transported pollutants as a quantitative value, while the experts are also interested in breaking down the value to find the actual sources of air pollution (e.g., vehicles or factories). Source apportionment methods [7] can be incorporated in the future for such functionality.

Design lessons. We conclude our design lessons learned from the presentation of massive propagation networks. To address the scalability issue that impedes the perception of significant patterns therein, we organize the propagation graphs in a stratified way that utilizes their innate hierarchy of common topology, frequent patterns, and instances, and design visualizations accordingly to support leveled exploration that imposes less cognitive load on users. Moreover, we visualize the frequent motifs and patterns with carefully designed glyphs that compactly summarize their attributes, empowering users to effectively analyze data from multiple perspectives. The detailed instances are expanded on demand. This approach coincides in spirit with the visual information-seeking mantra [51] and once again confirms its effectiveness in guiding the exploration of complex datasets.

8 Conclusion

In this study, we propose a novel visual analytics approach to incorporate domain knowledge in analyzing the uncertain propagation of air pollution. By closely collaborating with the experts, we characterized the user requirements in the topology-driven analysis of propagation processes and derived a set of extensive design goals accordingly to guide the subsequent visual design. Based on the requirements and design goals, we developed AirVis, a visual analytics system that assists users in hierarchically exploring and interpreting massive propagation patterns extracted with a novel FSM-based pattern mining framework. The effectiveness of our system is demonstrated via two case studies conducted on the real-world dataset and the positive feedback received from the experts. Our approach is also generalizable to other similar problems that involve the topology-driven analysis of massive small graphs. In the future, we would like to incorporate fine-grained atmospheric data to obtain more accurate results and deploy our system in the field to reveal the insights in pollution propagation to a wider audience, including environmental scientists and government officials.

Acknowledgments

We thank all reviewers for their constructive comments. The work was supported by National Key R&D Program of China (2018YFB1004300), NSFC-Zhejiang Joint Fund for the Integration of Industrialization and Informatization (U1609217), NSFC (61761136020), Zhejiang Provincial Natural Science Foundation (LR18F020001) and the 100 Talents Program of Zhejiang University.

REFERENCES

[1] G. L. Andrienko, N. V. Andrienko, P. Bak, D. Keim, and S. Wrobel. *Visual Analytics of Movement*. Springer, 2013.

[2] G. L. Andrienko, N. V. Andrienko, C. Hurter, S. Rinzivillo, and S. Wrobel. Scalable analysis of movement data for extracting and exploring significant places. *IEEE TVCG*, 19(7):pp–1078, 2013.

[3] N. V. Andrienko, G. L. Andrienko, L. Barrett, M. Dostie, and P. Henzi. Space transformation for understanding group movement. *IEEE TVCG*, 19(12):2169–2178, 2013.

[4] F. Beck, M. Burch, S. Diehl, and D. Weiskopf. A taxonomy and survey of dynamic graph visualization. *CGF*, 36(1):133–159, 2017.

[5] R. Borgo, J. Kehrer, D. H. S. Chung, E. Maguire, R. S. Laramee, H. Hauser, M. Ward, and M. Chen. Glyph-based visualization: Foundations, design guidelines, techniques and applications. In *Proc. of Eurographics, STARs*, pages 39–63, 2013.

[6] K. M. Borgwardt, H.-P. Kriegel, and P. Wackersreuther. Pattern mining in frequent dynamic subgraphs. In *Proc. of IEEE ICDM*, pages 818–822, 2006.

[7] M. Bove, P. Brotto, F. Cassola, E. Cuccia, D. Massabò, A. Mazzino, A. Piazzalunga, and P. Prati. An integrated PM2.5 source apportionment study: positive matrix factorisation vs. the chemical transport model camx. *Atmospheric environment*, 94:274–286, 2014.

[8] K. Buchin, B. Speckmann, and K. Verbeek. Flow map layout via spiral trees. *IEEE TVCG*, 17(12):2536–2544, 2011.

[9] M. Burch and S. Diehl. TimeRadarTrees: Visualizing dynamic compound digraphs. *CGF*, 27(3):823–830, 2008.

[10] D. Byun and K. L. Schere. Review of the governing equations, computational algorithms, and other components of the models-3 community multiscale air quality (cmaq) modeling system. *Applied mechanics reviews*, 59(2):51–77, 2006.

[11] G. Y.-Y. Chan, P. Xu, Z. Dai, and L. Ren. Vibr: Visualizing bipartite relations at scale with the minimum description length principle. *IEEE TVCG*, 25(1):321–330, 2019.

[12] S. Chen, X. Yuan, Z. Wang, C. Guo, J. Liang, Z. Wang, X. L. Zhang, and J. Zhang. Interactive visual discovering of movement patterns from sparsely sampled geo-tagged social media data. *IEEE TVCG*, 22(1):270–279, 2016.

[13] Y. Chen, P. Xu, and L. Ren. Sequence synopsis: Optimize visual summary of temporal event data. *IEEE TVCG*, 24(1):45–55, 2018.

[14] A. Daly and P. Zannetti. Air pollution modeling–an overview. *Ambient air pollution*, pages 15–28, 2007.

[15] M. Deshpande, M. Kuramochi, N. Wale, and G. Karypis. Frequent substructure-based approaches for classifying chemical compounds. *IEEE TKDE*, 17(8):1036–1050, 2005.

[16] A. Donnelly, B. Misstear, and B. Broderick. Real time air quality forecasting using integrated parametric and non-parametric regression techniques. *Atmospheric Environment*, 103:53–65, 2015.

[17] N. Ferreira, J. Poco, H. T. Vo, J. Freire, and C. T. Silva. Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips. *IEEE TVCG*, 19(12):2149–2158, 2013.

[18] M. Freire, C. Plaisant, B. Shneiderman, and J. Golbeck. Manynets: an interface for multiple network analysis and visualization. In *Proc. of ACM SIGCHI*, pages 213–222, 2010.

[19] C. W. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438, 1969.

[20] M. Greilich, M. Burch, and S. Diehl. Visualizing the evolution of compound digraphs with TimeArcTrees. *CGF*, 28(3):975–982, 2009.

[21] D. Guo. Visual analytics of spatial interaction patterns for pandemic decision support. *International Journal of Geographical Information Science*, 21(8):859–877, 2007.

[22] D. Guo and X. Zhu. Origin-destination flow data smoothing and mapping. *IEEE TVCG*, 20(12):2043–2052, 2014.

[23] S. Hadlak, H. Schumann, and H. Schulz. A survey of multi-faceted graph visualization. In *Proc. of EuroVis, STARs*, pages 1–20, 2015.

[24] X. Huang, Y. Zhao, C. Ma, J. Yang, X. Ye, and C. Zhang. Trajgraph: A graph-based visual analytics approach to studying urban network centralities using taxi trajectory data. *IEEE TVCG*, 22(1):160–169, 2016.

[25] A. Inokuchi, T. Washio, and H. Motoda. An apriori-based algorithm for mining frequent substructures from graph data. In *Proc. of PKDD*, pages 13–23, 2000.

[26] C. Jiang, F. Coenen, and M. Zito. A survey of frequent subgraph mining algorithms. *The Knowledge Engineering Review*, 28(1):75–105, 2013.

[27] L. Jiang and L. Bai. Spatio-temporal characteristics of urban air pollutions and their causal relationships: Evidence from beijing and its neighboring cities. *Scientific reports*, 8(1):1279, 2018.

[28] S. Kairam, D. L. MacLean, M. Savva, and J. Heer. Graphprism: compact visualization of network structure. In *Proc. of ACM AVI*, pages 498–505, 2012.

[29] M. Kampa and E. Castanas. Human health effects of air pollution. *Environmental pollution*, 151(2):362–367, 2008.

[30] D. Koracin, R. Vellore, D. H. Lowenthal, J. G. Watson, J. Koracin, T. McCord, D. W. DuBois, L.-W. A. Chen, N. Kumar, E. M. Knipping, et al. Regional source identification using lagrangian stochastic particle dispersion and hysplit backward-trajectory models. *Journal of the Air & Waste Management Association*, 61(6):660–672, 2011.

[31] M. Krzywinski, I. Birol, S. J. Jones, and M. A. Marra. Hive plots—rational approach to visualizing networks. *Briefings in bioinformatics*, 13(5):627–644, 2011.

[32] P. Li, R. Yan, S. Yu, S. Wang, W. Liu, and H. Bao. Reinstate regional transport of PM2.5 as a major cause of severe haze in beijing. In *Proceedings of the National Academy of Sciences*, volume 112, pages E2739–E2740. National Acad Sciences, 2015.

[33] X. Li, Y. Cheng, G. Cong, and L. Chen. Discovering pollution sources and propagation patterns in urban area. In *Proc. of ACM SIGKDD*, pages 1863–1872, 2017.

[34] Y. Li, J. Huang, and J. Luo. Using user generated online photos to estimate and monitor air pollution in major cities. In *Proc. of ACM ICIMCS*, page 79, 2015.

[35] W. Lin, Y. Zhou, H. Xu, J. Yan, M. Xu, J. Wu, and Z. Liu. A tube-and-droplet-based approach for representing and analyzing motion trajectories. *IEEE TPAMI.*, 39(8):1489–1503, 2017.

[36] D. Liu, D. Weng, Y. Li, J. Bao, Y. Zheng, H. Qu, and Y. Wu. Smartadp: Visual analytics of large-scale taxi trajectories for selecting billboard locations. *IEEE TVCG.*, 23(1):1–10, 2017.

[37] L. v. d. Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[38] H. McGowan and A. Clark. Identification of dust transport pathways from lake eyre, australia using hysplit. *Atmospheric Environment*, 42(29):6915–6925, 2008.

[39] S. Mei, H. Li, J. Fan, X. Zhu, and C. R. Dyer. Inferring air pollution by sniffing social media. In *Proc. of IEEE/ACM ASONAM*, pages 534–539, 2014.

[40] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[41] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Proc. of NIPS*, pages 3111–3119, 2013.

[42] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

[43] S. Navlakha, R. Rastogi, and N. Shrivastava. Graph summarization with bounded error. In *Proc. of ACM SIGMOD*, pages 419–432, 2008.

[44] C. Nobre, M. Streit, and A. Lex. Juniper: A tree+table approach to multivariate graph visualization. *IEEE TVCG.*, 25(1):544–554, 2019.

[45] H. Qu, W.-Y. Chan, A. Xu, K.-L. Chung, K.-H. Lau, and P. Guo. Visual analysis of the air pollution problem in hong kong. *IEEE TVCG*, 13(6):1408–1415, 2007.

[46] J. Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

[47] R. Scheepens, N. Willems, H. Van de Wetering, G. Andrienko, N. Andrienko, and J. J. Van Wijk. Composite density maps for multivariate trajectories. *IEEE TVCG*, (12):2518–2527, 2011.

[48] C. Schulz, A. Nocaj, J. Görtler, O. Deussen, U. Brandes, and D. Weiskopf. Probabilistic graph layout for uncertain network visualization. *IEEE TVCG.*, 23(1):531–540, 2017.

[49] M. Sedlmair, M. D. Meyer, and T. Munzner. Design study methodology: Reflections from the trenches and the stacks. *IEEE TVCG.*, 18(12):2431–2440, 2012.

[50] J. H. Seinfeld. Urban air pollution: state of the science. *Science*, 243(4892):745–752, 1989.

[51] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the 1996 IEEE Symposium on Visual Languages*, pages 336–343, 1996.

[52] A. Stein, R. R. Draxler, G. D. Rolph, B. J. Stunder, M. Cohen, and F. Ngan. Noaa's hysplit atmospheric transport and dispersion modeling system. *Bulletin of the American Meteorological Society*, 96(12):2059–2077, 2015.

[53] Y. Sun, Z. Wang, O. Wild, W. Xu, C. Chen, P. Fu, W. Du, L. Zhou, Q. Zhang, T. Han, Q. Wang, X. Pan, H. Zheng, J. Li, X. Guo, J. Liu,

and D. R. Worsnop. “APEC Blue”: Secondary Aerosol Reductions from Emission Controls in Beijing. *Scientific Reports*, 6:20668, Feb. 2016.

[54] The World Air Quality Project. Air Pollution in Asia: Real-time Air Quality Index Visual Map. https://aqicn.org/map/, 2019. [Online; accessed 30-Mar-2019].

[55] C. Tominski, H. Schumann, G. Andrienko, and N. Andrienko. Stacking-based visualization of trajectory attribute data. *IEEE TVCG*, 18(12):2565–2574, 2012.

[56] S. van den Elzen, D. Holten, J. Blaas, and J. J. van Wijk. Dynamic network visualization withextended massive sequence views. *IEEE TVCG*, 20(8):1087–1099, 2014.

[57] N. Vanetik, S. E. Shimony, and E. Gudes. Support measures for graph data. *Data Mining and Knowledge Discovery*, 13(2):243–260, 2006.

[58] T. Von Landesberger, F. Brodkorb, P. Roskosch, N. Andrienko, G. Andrienko, and A. Kerren. Mobilitygraphs: Visual analysis of mass mobility dynamics via spatio-temporal graphs and clustering. *IEEE TVCG*, 22(1):11–20, 2016.

[59] T. von Landesberger, A. Kuijper, T. Schreck, J. Kohlhammer, J. J. van Wijk, J. Fekete, and D. W. Fellner. Visual analysis of large graphs: State-of-the-art and future research challenges. *CGF*, 30(6):1719–1749, 2011.

[60] L. Wang, Z. Liu, Y. Sun, D. Ji, and Y. Wang. Long-range transport and regional sources of PM2.5 in beijing based on long-term observations from 2005 to 2010. *Atmospheric Research*, 157:37–48, 2015.

[61] Y. Wang, X. Zhang, and R. R. Draxler. Trajstat: Gis-based software that uses various trajectory statistical analysis methods to identify potential sources from long-term air pollution measurement data. *Environmental Modelling and Software*, 24(8):938–939, 2009.

[62] W. Wu, J. Xu, H. Zeng, Y. Zheng, H. Qu, B. Ni, M. Yuan, and L. M. Ni. Telcovis: Visual exploration of co-occurrence in urban human mobility based on telco data. *IEEE TVCG*, 22(1):935–944, 2016.

[63] Y.-F. Xing, Y.-H. Xu, M.-H. Shi, and Y.-X. Lian. The impact of PM2.5 on the human respiratory system. *Journal of thoracic disease*, 8(1):E69, 2016.

[64] X. Yan and J. Han. gSpan: Graph-based substructure pattern mining. In *Proc. of IEEE ICDM*, pages 721–724, 2002.

[65] Y. Yang, T. Dwyer, S. Goodwin, and K. Marriott. Many-to-many geographically-embedded flow visualisation: an evaluation. *IEEE TVCG*, 23(1):411–420, 2017.

[66] X. Yi, J. Zhang, Z. Wang, T. Li, and Y. Zheng. Deep distributed fusion network for air quality prediction. In *Proc. of ACM SIGKDD*, pages 965–973, 2018.

[67] V. Yoghourdjian, T. Dwyer, K. Klein, K. Marriott, and M. Wybrow. Graph thumbnails: Identifying and comparing multiple graphs at a glance. *IEEE TVCG*, (1):1–1, 2018.

[68] D. Zhang, J. Liu, and B. Li. Tackling air pollution in China—what do we learn from the great smog of 1950s in london. *Sustainability*, 6(8):5322–5338, 2014.

[69] J. P. Zhang, T. Zhu, Q. Zhang, C. Li, H. Shu, Y. Ying, Z. Dai, X. Wang, X. Liu, A. Liang, et al. The impact of circulation patterns on regional transport pathways and air quality over beijing and its surroundings. *Atmospheric Chemistry and Physics*, 12(11):5031–5053, 2012.

[70] S. Zhang, G. Jin, X.-S. Zhang, and L. Chen. Discovering functions and revealing mechanisms at molecular level from biological networks. *Proteomics*, 7(16):2856–2869, 2007.

[71] Y. Zheng. *Urban Computing*. MIT Press, 2019.

[72] Y. Zheng, X. Chen, Q. Jin, Y. Chen, X. Qu, X. Liu, E. Chang, W.-Y. Ma, Y. Rui, and W. Sun. A cloud-based knowledge discovery system for monitoring fine-grained air quality. *preparation, Microsoft Tech Report, http://research. microsoft. com/apps/pubs/default. aspx*, 2014.

[73] Y. Zheng, F. Liu, and H.-P. Hsieh. U-air: When urban air quality inference meets big data. In *Proc. of ACM SIGKDD*, pages 1436–1444, 2013.

[74] Y. Zheng, X. Yi, M. Li, R. Li, Z. Shan, E. Chang, and T. Li. Forecasting fine-grained air quality based on big data. In *Proc. of ACM SIGKDD*, pages 2267–2276, 2015.

[75] Z. Zhou, L. Meng, C. Tang, Y. Zhao, Z. Guo, M. Hu, and W. Chen. Visual abstraction of large scale geospatial origin-destination movement data. *IEEE TVCG*, 25(1):43–53, 2019.

[76] J. Y. Zhu, C. Zhang, H. Zhang, S. Zhi, V. O. Li, J. Han, and Y. Zheng. pg-causality: Identifying spatiotemporal causal pathways for air pollutants with urban big data. *IEEE TBD*, 4(4):571–585, 2018.